Information Research - Vol. 31 No. iConf (2026)

Reframing data papers as boundary objects: aligning data narratives with reuse-oriented user expectations

Cuiyu Qin, Qingyu Duan, Lei Xu, and Xiaoguang Wang

DOI: *https://doi.org/10.47989/ir31iConf64123*

Abstract

Introduction. This study investigates how the narrative structure of data papers, as boundary objects, aligns with the needs of data reusers, focusing on both document structure and reusers' preferences.

Method. Two approaches were employed: (1) a textual analysis of discourse components and data events, and (2) an exploration of reusers' preferences and perceived use value through interviews.

Analysis. A total of 210 data papers were randomly sampled for content analysis. Interview data were analysed using both inductive and deductive thematic analysis.

Results. Data papers share several discourse components with traditional academic papers but also include unique components such as Data Value (V), Usage Notes (U), Data Availability (A), and Quality Control (Q). A total of 18 types of data events were identified. Data reusers' needs were categorised into five dimensions: data collection, content, structure, analysis, and reuse. Reusers perceive the functional, social, and cognitive value of data papers through research actions.

Conclusion(s). Data papers meet foundational needs in discovery, filtering, and comprehension, but fail to address application-level needs like trust assessment, accessibility, and reconstruction hints. This study also indicates the importance of data papers' linking with data repositories and corresponding academic papers.

Introduction

Open science has reached a new stage of global consensus. According to international scientific data repository platforms such as re3data.org and FAIRsharing, over 4,200 platforms worldwide now offer data publishing and utilisation services, contributing to the gradual formation of an open data innovation ecosystem. While significant achievements have been made in the development of scientific data publishing, the issue of data reuse post-publication has emerged as a widely debated topic, particularly regarding how to unlock the value of data and harness its potential as a critical component of scientific research (Borgman & Groth, 2025). Data papers, a novel genre within scholarly publishing driven by the open-science movement, facilitate data discovery and reuse through detailed descriptions of research datasets (Callaghan et al, 2012; Candela et al, 2015). In addition to the term Data Paper, various journals and publishing platforms employ terms such as *Data Article*, *Data Descriptors*, *Data in Brief*, *Data Note*, *Dataset Paper*, and *Data Original Article*, which are synonymous or closely related. Through establishing standardised, citable, and evaluable frameworks (Penev et al., 2012; Gregory et al., 2023), data papers create shared reference points among research communities, thereby exhibiting characteristics of boundary objects (Østerlund & Crowston, 2019).

In practical scientific communication, however, there remain many challenges regarding whether data papers effectively meet the needs of data reusers. Currently, most data papers primarily describe information about data production, with limited attention to the general attributes of datasets (Kim, 2020). Data papers should provide sufficient information to enable potential researchers to overcome data isolation issues during reuse. (Chao, 2015; Edwards et al., 2011). Genre analysis of discourse can enhance researchers' comprehension of the form, content, and purpose of texts (Swales, 2004; Zhang et al., 2010). Therefore, this paper adopts a dual perspective, focusing on both text and reusers, to investigate how the narrative structure of data papers, as boundary objects, aligns with the needs of data reusers. This alignment aims to enhance the usability and value of data within broader research practices.

Literature review

Data paper as boundary objects

Star and Griesemer (1989) introduced the concept of Boundary Objects (BOs) in their study of information practices at Berkeley's Museum of Vertebrate Zoology. They defined BOs as objects, documents, tools, or concepts that serve as bridges between different communities, acting as translation mechanisms that help maintain coherence, facilitate communication, and support collaboration. They categorised BOs into four types: repositories, ideal types, coincident boundaries, and standardised forms (Star & Griesemer, 1989). The concept of boundary objects is highly relevant to the field of documentation theory (Björk, 2015; Huvila, 2011), as documents play a dual role as both tangible, visible entities, and as transparent, infrastructural mediators (Boell & Hoof, 2015). From the perspective of documentation, one significant aspect of boundary objects is that they help clarify the functions of documents and other information objects within various social or individual contexts (Frohmann, 2004; Yeo, 2008). From the perspective of boundary objects, data papers can be viewed as a document type that bridges the roles of data creators, data reusers, and data publishers, facilitating the discovery, sharing, and reuse of research data. Data are purposeful collections of facts, observations, or objects used as evidence for research (Borgman, 2015; Zins, 2007). Thorough documentation of data collection and preparation can enhance user trust and improve data quality (Das, 2021; Moher et al., 2020).

Narrative structure of data papers

Scientific discourse can be viewed as a form of narrative, specifically, a scientific narrative (Sheehan & Rode, 1999), which is utilised for the interpretation of science (Yeo & Gilbert, 2014). In data papers, the narrative structure serves as the overall framework for information dissemination, integrating discourse structure, which refers to the organisation of the text, with data events, which describe activities such as data generation, processing, and application. These two elements complement each other, collectively constructing the narrative structure of the data paper and ensuring effective academic communication. Existing studies primarily analyse the content-structure of data papers from a document perspective, drawing on theories such as genre theory and discourse analysis. One approach involves analysing data policies and submission guidelines from various journals to identify common content elements and discourse features (Candela et al., 2015; Roa-Martinez et al., 2017). A second approach examines data papers published in specialised journals. For instance, Li and Jiao (2022) performed a brief rhetorical move analysis on the abstracts of 360 papers sampled from *Scientific Data* and *Data in Brief*. Their findings indicate that the abstracts of data papers incorporate a combination of IMRaD and data-oriented rhetorical moves.

Recently, some studies related to data events have emerged from the perspectives of data management practices. Data Practices and Curation Vocabulary (DPCVocab) specifies the relationships among data practices in research, types of data produced and used, and curation roles and activities (Chao et al., 2015). Li et al. (2020) identified 17 types of data events in data papers from biodiversity. However, the distribution of data events within the discourse structure of data papers requires further systematic exploration. These studies further prompt us to explore some critical dimensions such as the granularity of data event narration in data papers and potential disciplinary variations in the representation of data events.

Additionally, other artifacts such as paradata and data provenance models serve functions similar to data events in data papers. Paradata (Data on the process of its creation, curation, and use) can improve the usefulness of research data (Huvila, 2022; Huvila et al., 2025). While Paradata focuses on detailed records of data collection and processing, data papers not only narrate the process of data but also describe its content. The Open Provenance Model and the PROV provenance standard (Moreau et al., 2011) formalise the description of elements such as actions, time, place, and objects of data events, as well as the interrelationships among events, ensuring the accurate recording of scientific data.

Data reuse

Data reuse behavior parallels information behavior in tasks and activities such as need identification, retrieval, selection, acquisition, and utilisation (Wang et al., 2021). Scientific data reuse requires not only an understanding of the data content but also a comprehensive understanding of how the data was created, how it was previously processed, interpreted, and used (Faniel & Yakel, 2017; Voss, 2012). Reusers evaluate the usability of data using information such as collection methods, instrumentation, observational conditions, provenance, and data processing (Gregory et al., 2019). Specific information structures attached to datasets can also facilitate certain sense-making patterns (Koesten et al., 2021). Candela et al. (2011) mentioned that descriptions of reuse information are often neglected by data journals. There is often a difference in opinion between data creators and users about which process data is necessary (Börjesson et al., 2022). This noticeable gap in what creators and users consider important makes it challenging to create and provide process data that is meaningful to both parties.

Existing literature either focuses on data description and data policies from a textual perspective or analyses data reuse practices from the perspective of reusers. However, the alignment between the narrative structure of data papers and the needs of data reusers has not been sufficiently addressed. Therefore, this paper analyses the gap between the narrative structure of data papers and the needs of reusers and explores how optimising the narrative structure can strengthen its function as a boundary object.

Method

To explore the alignment between the narrative structure of data papers as boundary objects and the needs of data reusers, this study approaches the issue from two perspectives: (1) an analysis of the narrative structure of data papers from a textual perspective, and (2) an investigation of the preferences and perceived use value of data papers from the perspective of data reusers. The research framework is presented in Figure 1. The analysis from the textual perspective reveals the descriptive patterns of components such as data generation, processing, and application in data papers, providing a theoretical basis for the design of subsequent interviews. There are differences in the cognitive understanding of the required process data between data creators and reusers (Börjesson et al., 2022). Based on this, the interviews from the data reusers' perspective further explore their needs and experiences with various components of data papers, particularly their expectations regarding data processing details, providing specific directions for improving the textual narrative structure of data papers.

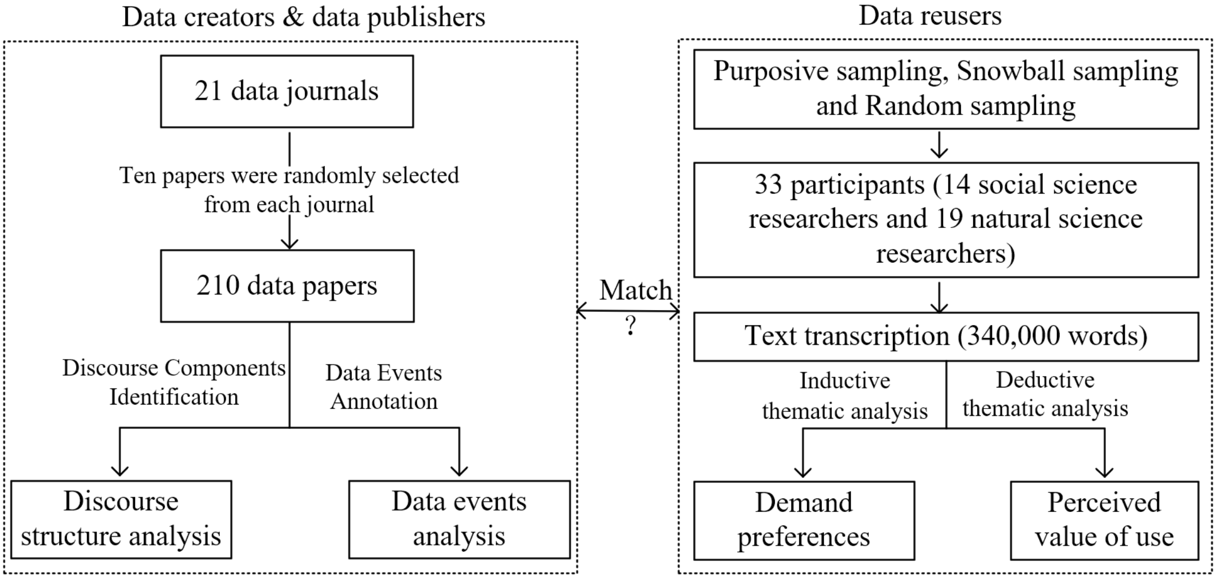
Data creators & data publishers
21 data journals
Ten papers were randomly selected from each journal
210 data papers
Discourse Components Identification
Data Events Annotation
Discourse structure analysis
Data events analysis
Match ?
Data reusers
Purposive sampling, Snowball sampling and Random sampling
33 participants (14 social science researchers and 19 natural science researchers)
Text transcription (340,000 words)
Inductive thematic analysis
Deductive thematic analysis
Demand preferences
Perceived value of use

Figure 1. Research framework

Textual analysis of data papers

In 2019, Schöpfel et al. compiled a comprehensive list of 82 data journals, including 28 exclusively dedicated to data. This study reviewed the operational status of these journals, excluding those that ceased publication, and updated the list to include previously omitted journals through online resources, literature, and relevant reports. The final sample consists of 21 active data journals from prominent publishers such as Elsevier, Springer Nature, and John Wiley, along with China Scientific Data, the only Chinese academic journal focused on scientific data across disciplines. A detailed list of the journals is provided in Table 1. By examining journal websites and published papers, this study assessed the subject and thematic scope of each journal. The 21 data journals span disciplines including physical chemistry, computer science, biomedical sciences, geography, agriculture, humanities, social sciences, and multidisciplinary fields. By reviewing the official websites of the data journals, we confirmed that the documents were data papers. This study randomly selected ten data papers published between 2021 and 2022 from each of these journals, yielding a total of 210 samples. Due to limited paper availability, the *OHD* journal contributed papers from 2015-2022, while the OJB journal contributed papers from 2020-2021, as 2022 papers were not available at the start of the study (September 2022).

Data journal title	Abbreviation	Publishers	Subject Scope
Atomic Data and Nuclear Data Tables	ADNDT	Elsevier	Physics
Chemical Data Collections	CDC		Chemical Engineering
Data in Brief	DIB		Comprehensive Subject
BMC Research Notes	BMC-RN	Springer Nature	Biology, Medical Sciences
Scientific Data	SD		Comprehensive Subject
Geoscience Data Journal	GDJ	John Wiley	Geography
The International Journal of Robotics Research	IJRR	Sage	Computer Applications
Biodiversity Data Journal	BDJ	Pensoft Publishers	Biology
Open Health Data	OHD	Ubiquity Press	Medical Sciences
Open Journal of Bioresources	OJB		Biology
Journal of Open Psychology Data	JOPD		Psychology
Data	Data	MDPI	Comprehensive Subject
Research Data Journal for the Humanities and Social Sciences	RDJHS	Publons	Social Sciences
Earth System Science Data	ESSD	Copernicus Publications	Earth Sciences
GigaScience	GS	Oxford Academic	Biology
Internet Archaeology	IA	University of York	Archaeology
Journal of Chemical & Engineering Data	JCED	American Chemical Society	Chemistry, Engineering
Journal of Physical and Chemical Reference Data	JPCRD	AIP	Chemistry, Physics
Open Archaeology	OA	Sciendo	Archaeology
Open Data Journal for Agricultural Research	ODJAR	Wageningen UR*Alterra	Agriculture
China Scientific Data	CSDATA	Chinese Academy of Sciences	Comprehensive Subject

Table 1. Data journal samples and their basic information.

The narrative structure of data papers was analysed by extracting discourse components, such as chapter titles, from the 210 selected papers, using associated writing templates and guidelines. Two coders engaged in iterative discussions to resolve ambiguities in the identification process. Discourse components were categorised and labelled accordingly.

For the analysis of data events, 210 papers were inductively coded using NVivo12 software to explore the types and functions of data events. Data life cycle models (Plale & Kouper, 2017), DPCVocab (Chao et al., 2015), and Li et al. (2020) for biodiversity data events were referenced to ensure comprehensive coverage of core data events. Initial coding of 63 papers resulted in the identification of data events, followed by iterative coding of the remaining 147 papers until saturation was reached. Expert consultations led to the consolidation of coding nodes, resulting in 18 types of data events. To verify the consistency of coding, independent coding was performed by students from the information management and chemistry programs, yielding a Kappa coefficient greater than 0.85, confirming high reliability.

User interviews

A combination of purposive sampling, snowball sampling, and random sampling was employed to recruit researchers with experience in data reuse. All participants had previously published research based on data. After recruiting 33 participants, theoretical saturation was reached, with each interview lasting approximately 36 to 84 minutes. Early-career researchers, with limited resources, rely heavily on public data and existing findings, making data acquisition, processing, and reuse crucial (Campbell et al., 2019). Moreover, early-career researchers exhibit a high level of acceptance of emerging technologies and methods (Nicholas et al., 2018), demonstrating flexibility and a willingness to explore various data reuse pathways. Future research could expand the sample size to include more senior researchers' perspectives to improve the broader applicability of the results. The participants included 14 social science researchers and 19 natural science researchers. The sample included 23 doctoral students, 7 master's students, and 3 postdoctoral researchers or lecturers, with approximately 340,000 words of interview transcripts. An inductive and deductive thematic analysis approach was employed. Two doctoral students initially coded five interview transcripts collaboratively to construct a coding book with a coding consistency of 0.811. The first author then completed the coding analysis of the remaining transcripts. The interview outline is as follows:

- What information do you focus on when reading data papers? Do data papers provide sufficient information to support data reuse?
- How do you assess data reliability?
- What role do you believe data papers play in your research?

Form and function of data papers: a textual perspective

Discourse components

Data papers share several discourse components with traditional academic papers, such as the Introduction (I), Materials and Methods (M), Data Description (De), Results (R), Discussion (D), and Conclusion (C). Data papers also include unique components, such as Data Value (V), Usage Notes (U), Data Availability (A), and Quality Control (Q). These components provide readers with the essential contextual information for understanding the data. Among these, Materials and Methods (M) and Data Description (De) are two crucial discourse components in data papers, as they enable readers to comprehend generation methods and scientific data comprehensively, facilitating the discovery, exploration, evaluation, and reuse of scientific data.

Data events in data papers

This study presents a series of events related to these data, categorises their types, and analyses their functions. In total, 2595 sentences containing data events were identified in the sample. Among these sentences, 3028 data event instances were identified. Based on this sample, our final data-event classification scheme is presented in Table 2, with examples for each type.

Data event	Definition	Example
Data collection (DCll)	The activities of collecting data in a variety of ways, include retrieving data from sources such as databases and search engines, surveying data with a tool, and designing experiments to obtain data, etc.	Data from the daily diary study **was collected** between October 2016 and February 2017. (12-1)
Data classification (DCl)	The activities of classifying data into a category or dividing a dataset into subcategories based on a criterion	The data subsets **are divided** by site and treatment of the Small digestate experiment. (15-1)
Data extraction (DExt)	The activities of extracting data of a certain type or attribute from datasets for display or further analysis to obtain new data.	From 13 to 54 data **were selected** for the 20 esters classified with code C in Table 2. (13-6)
Digitisation (Dig)	The activities of scanning the images and manuscripts and converting them into digital formats or digitally drawing objects or relics.	All three floors which cover an area of over 15,800 m2 **are scanned** using a Navvis M6 device shown in Figure 1. (20-6)
Data labeling and coding (DL&C)	The activities of adding labels, annotations, and encodings to data.	As shown in Table 1, the code S **was assigned** to 42 esters. (13-6)
Data analysis (DAn)	The activities of analysis of collected data using statistical or analytical procedures, calculation of data using formula algorithms, or comparison of two or more data to obtain new information, etc.	The measured data **were compared with** the literature data, and the deviations **were analysed**. (11-7)
Data linking and integration (DL&I)	The activities of associating data from different sources with related fields or attributes and adding data to the original dataset.	Similarly, this **was merged with** the school holidays data to identify whether the day was a school holiday or not. (19-10)
Data removal (DRem)	The activities of deleting and filtering data from original dataset.	Data with significant observed sensor error **was removed** during processing. (3-9)
Data modification (DMod)	The activities of modification, replacement, and updating of the original data.	Data **were corrected** and **modified** only to the extent necessary to improve clarity and usability. (2-1)
Data list (DList)	The activities of listing and displaying data in text, tables, etc.	The data resources that are used and linked in GEMI **are listed** below. (8-5)
Data visualisation (DVis)	The activities of representing data in a visual way.	Examples of the paper tables with the observations **are presented** in Figure 3. (8-7)
Data Desensitisation (DDes)	The activities of anonymisation and de-privacy of data.	However, these data **were anonymised** to prevent individual respondent identification. (5-2)
Data Standardisation (DSta)	The activities of normalising data structures according to a standard	The data **is organised** in accordance with the Brain Imaging Data Structure (BIDS) specification version 1.5.0. (19-2)
Data identification (DIde)	The activities of assigning identifiers to the data.	Similarly, each unique literature source and author **are assigned** a unique identifier in the Reference Table (refID) and Author Table (auID). (13-7)

Data formatting (DFor)	The activities of conversing data (file) format to another.	We have done that by **converting** the Excel files to csv-text files. (18-7)
Data validation (DVal)	The activities of controlling the data quality, including checking and correcting data errors, data biases, etc.	The integrity of all the data **was checked** before sharing this dataset. (19-5)
Data storage (DSto)	The activities of recording, saving, and storing data.	The weather data were observed every 10s, and 15-minute averages **were recorded and saved**. (15-2)
Data publication (DPub)	The activities such as uploading data to data repository, submitting it to data centre or agency department, or making it publicly available	Data **were uploaded** to the Open Science Framework by October 10, 2017. (12-1)

Table 2. Types of data events.

Statistical analysis of data events

Figure 2 illustrates the total frequency of each data event category at the sentence and paper levels. Among these, data collection (DCll), data list (DList), data analysis (DAn), data visualisation (DVis), data storage (DSto), and data validation (DVal) were the most frequently described events in data paper narrative process, and the sum of these six data event instances (2235) accounts for more than 70%.

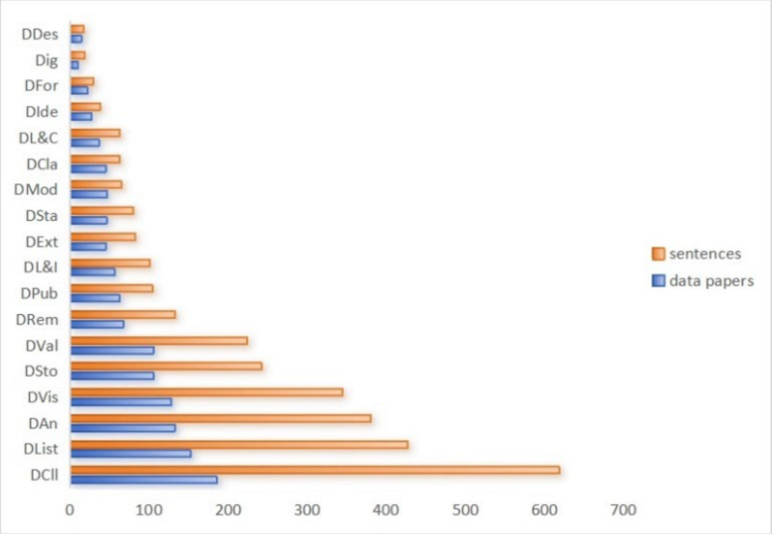

Figure 2. Frequency of each data event.

Distribution of data events across fields

As shown in Figure 3, there are certain differences in the distribution of data events among different fields, with the number of data events in the humanities and social sciences generally being lower than that of the other six disciplines, while geography and interdisciplinary data papers describe a higher number of data events compared to other disciplines. The distribution of data events varies greatly among the seven disciplines, with a certain type of event appearing more frequently in one discipline than in others. The reasons for the distribution differences in data events can largely be attributed to the characteristics of the data, theme orientation of data journals, and data publishing policies. For instance, in geography, the description of a data publication (DPub) event is far more prevalent than in the other four disciplines, excluding the interdisciplinary category. The *GDJ* established an Open Data badge to reward the open sharing of scientific data, which facilitates the reproducibility of scientific results. In humanities and social sciences, data formatting (DFor) and digitisation (Dig) events rank third. The *JOPD* journals in the field of psychology require data to be provided in non-proprietary formats. In archaeology, data in journals such as IA and OA often involve digitisation processes, such as OCR recognition and image segmentation of ancient text images. The humanities and social sciences also have the highest number of data desensitisation (DDes) events, likely because of the nature of their data, which often involve survey and interview data obtained through field surveys, questionnaires, and field interviews. Such data frequently include personal privacy issues.

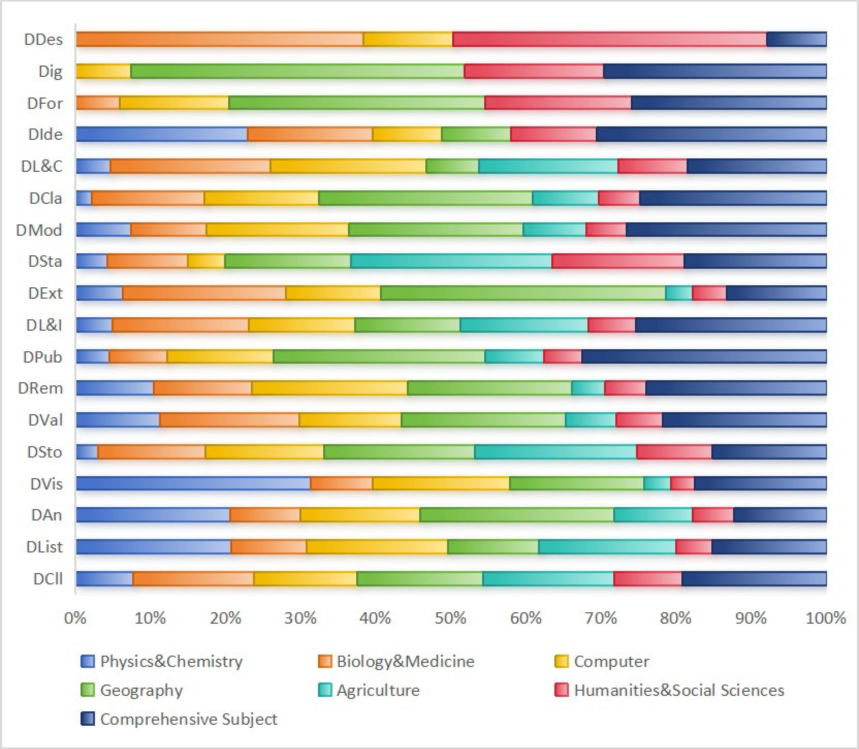

Figure 3. Proportion of 18 types of data events by subject topics.

Discourse components distribution of data events

The coverage of the 18 data events across the discourse components is shown in Figure 4. Data events were more frequently distributed in the Materials and Methods (M) and Data Description (De) components. This indicates that the primary function of these two components is to describe information about when, where, what, and how data were collected, as well as the processes of data modification, analysis, classification, integration, formatting, identification, storage, and publication.

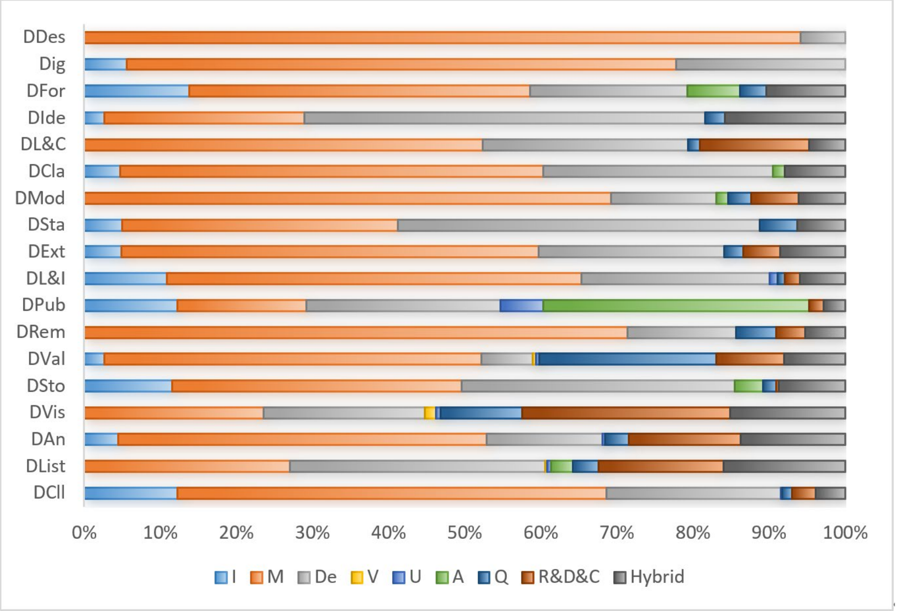

Figure 4. Distribution of 18 data events in discourse components.

Note: The narrative structure of the OA, *IA*, *JPCRD*, *ODJAR*, and *IJRR* journals varies significantly, with discourse components being highly variable and irregular. This variability may be determined by the disciplinary attributes of the data or by the lack of a mandatory writing policy or the absence of a provided writing template for data papers in the journals. This study treats these as hybrid components.

Although the proportion of data events in the hybrid component was relatively small, a diverse array of event types was still represented within this component. This suggests that even if some data papers do not follow a standardised discourse structure, the essential functions of describing and documenting the data formation process are maintained. Data publication (DPub) and data validation (DVal) events comprised significant proportions within the data accessibility (A) and quality control (Q) components. This is due to the role of these components in elucidating information related to the acquisition, usage, and verification of datasets. The distribution of data events within the discourse of data papers provides insights into the positional information of these events, facilitating strategic reading about data papers, and enabling readers to obtain contextual information about the data during their research.

Demand preferences and perceived use value for data papers: reuser's perspective

User preferences in data papers

Through in-depth interview analysis, data reusers' preferences for data papers can be categorised into five dimensions: data collection, content, structure, analysis, and reuse. Each dimension is further broken down into second-order and first-order themes, providing a systematic overview of reusers' needs, as shown in Table 3. In analysing the interview data, we focused on the specific needs expressed by data reusers, categorising their preferences and needs statistically, and mapping them to the components and events of data papers to identify gaps. Compared to the findings from textual analysis, new focal points emerged in the reusers' demands, including experimental phases, publication platform reputation, and reuse statistics. The most critical data description needs include the implementation process, analysis results, raw records, content attributes, and data distribution features. These needs align closely with the identified '*Materials and Methods*,' '*Data Description*' discourse components and data event.

However, there remain gaps between the reusers' needs and the information provided in data papers. First, regarding data collection dimensions, data papers seldom specify concrete experimental phases. Data papers frequently focus on describing the final dataset, with less coverage of the various stages in the data generation process. For example, in medical experiments, data typically undergoes multiple phases (such as Phase I, II, III, etc.), yet these phases are seldom mentioned in data papers, and reusers would prefer more explicit, actionable information. Second, as for data processing methods, although coding or transformation processes are described, the lack of transparency in code, scripts, and processing logic hinders the reusers' ability to directly transfer and reuse data. Third, in the data analysis dimension, while data papers generally document the analysis process (i.e., DAn events), the publication of corresponding results is often absent, limiting reusers' understanding of data interpretation and application. This also indicates the importance of linking data papers with their corresponding academic publications. Finally, regarding data reuse, data papers often lack cross-dataset comparisons or third-party validation of data quality, which are crucial for establishing data reliability. Reusers typically rely on data repositories for information on platform reputation and reuse metrics.

Aggregated Dimension	Second-order Themes	First-order Themes	Match	Gap
Data Collection	Collection Tools	Instruments, chemicals, software, surveys, interview guidelines	I; DCll, DExt, DCla, Dig, DL&C, DL&I, DRem, DMod, DDes, DSta, DFor, DIde, DVal	Data papers primarily focus on successful final data results, with limited description of process data (i.e., data from specific stages).
	Collection Methods	Experimental design, sampling methods, survey design, interview guidelines, program, and logic		
	Implementation Process	Collection location, collection time, collector, collection accuracy, collection site/band/energy range, collection process, experimental conditions/environment, *experimental stage		
	Data Processing Methods	Computation, processing programs, coding, annotations, transformation		
	Data Processing Effects	Data processing outcomes		
	Collection Background	Project description, funders, purpose of collection, subjects of collection and their conditions		
Data Content	Raw Records	Observations/records, main variables	De; DList, DVis	-
	Content Attributes	Name, subject, sample size/data size, original data source		
	Data Distribution Features	Time span, geographic coverage, type coverage, missing values, statistical description		
	Data File	Data file (and structure)		
	Content Explanation	Codebook, code explanation		
Data Structure	Data Format	Data size, format, version		
Data Analysis	Analysis Process	Methods, software, models, logic	DAn	Lack of publication output description
	Analysis Results	Results, publication output		
*Data Reuse	Scope of Use	Data usage, limitations	V, U; DSto, DPub	Not all data papers cover these components, and platform reputation requires reusers to obtain this information elsewhere
	Access Information	Data ID, *platform reputation, access level, code storage location		
	Data Quality	Comparisons with similar datasets, quality verification	Q; DVal	Data papers focus on self-validation, lacking cross-dataset comparison and third-party verification
	Reuse Metrics	Download count, reuse outcomes	-	Data papers often lack descriptions of reuse status, relying on data repositories for this information

Table 3. Data reusers' preferences for the needs of data papers.

Interview results indicate that researchers in natural sciences tend to clarify research object-related information from existing data, such as gene sequence alignment analysis (S2) ('S' is the abbreviation for Natural Science, and 'SS' is the abbreviation for Social Science), functional sites of gene activity (S18), and specific terrain and topography of species growth (S19). These reusers require more detailed information about the data production process. In contrast, researchers in the social sciences seek more comprehensive descriptions of data content, particularly content attributes and distribution features. They are also more inclined to use scientific data as prospective data, drawing new insights (SS7, SS8), developing new models, or using it as evidence to advance academic innovation. Furthermore, some researchers mentioned that when scientific data is used as background data, it is often explicitly stated in the results to indicate the rigor and standardisation of the research process (SS9).

Perceived use value of data papers

Data papers, as carriers of scientific data descriptions, serve as boundary objects within research data events, facilitating knowledge flow and coordinating actions across diverse research communities. Data reusers' requirements within data papers span dimensions such as data collection, content, structure, analysis, and reuse. These needs reflect practical demands arising from activities like data discovery, comprehension, and reconstruction. To better understand how these needs translate into perceived use value in research practice, interview transcripts were coded to identify the utilitarian value supporting specific research actions. The findings reveal that researchers perceive the functional, social, and cognitive value of data papers through concrete actions, as summarised in Table 4.

Aggregate Dimension	Second-Level Theme	First-Level Theme & Example
Functional Value (supporting scientific actions)	Data Discovery	'It essentially displays the major cities' nighttime lights, so we just need to check the year, and also the satellite model.' (S5, Discovery & Filtering Basis)
	Data Filtering	
	Data understanding	'[Some] it already plots the graph for you, so it's obvious, like it has distinct characteristics of clustered distribution.' (SS3, Understanding Basis)
	Trust Basis	'In high-energy nuclear physics, we basically consider this a relatively authoritative [data] website.' (S10, Trust Indicator)
	Access to Data	'At that time, downloading was troublesome, but later I asked my senior for the link and got it from him.' (S11, Access Path)
	Reconstruction Hints	'For example, I encountered two different annotation formats earlier, so when evaluating algorithms, I might need to write two different codes to extract some background facts.' (S11, Reconstruction hints)
Social Value (bridging research communities)	High recognition	'In high-energy nuclear physics, we basically consider this a relatively authoritative [data] website.' (S10, Source Recognition)
		'If it has more downloads, it probably means it's more authoritative, or it's been used by many people.' (SS8, Usage Recognition)
Cognitive Value (facilitating innovation and methodological adaptation)	Inspiring New Ideas	'It even has Japanese statistics... In our field, we have a specialised journal focused on Japanese issues. If I get this data, I could publish in it.' (SS9, Inspiring New Ideas)
	Acquiring New Methods	'If I see an accuracy of 92.4%, I won't continue with this research, but I'll learn from the method and see if I can apply it to other directions.' (S14, Acquiring New Methods)

Table 4. Perceived use value of data papers.

Functional value manifests in data papers effectively meeting foundational needs such as data discovery, filtering, and comprehension by providing detailed collection methods, data attributes, and analytical results. However, application-level needs such as trust assessment, accessibility, and reconstruction hints remain inadequately addressed. Reusers primarily rely on data repository platforms to obtain such information. This suggests that the functionality and value of data papers are significantly contingent upon their linkage with data repository platforms; the association between the two is essential for facilitating effective data reuse.

Social value manifests in the authority and recognition information conveyed by data papers. Researchers often regard metrics such as download counts, platform reputation, and source authority as trust indicators, thereby fostering consensus across different research communities and enhancing data dissemination and utilisation efficiency. Nevertheless, data papers remain deficient in conveying such social value information, particularly concerning platform reputation and reuse metrics.

Cognitive value manifests in how data papers' narrative frameworks and analytical processes inspire novel research concepts among researchers. They offer insights for method transfer or adaptation, enabling interdisciplinary innovation and collaboration. Yet, when contrasted with the textual analysis findings, the realisation of these values remains constrained. Current data papers focus predominantly on data collection and processing, with minimal attention given to how data can support interdisciplinary collaboration or facilitate the transfer and innovation of methods across diverse research contexts.

Discussion

Main findings

Role of data papers as boundary objects

Data papers possess functional value in supporting research activities, social value in bridging research communities, and cognitive value in fostering innovation and methodological reference. Data papers exhibit common features, with standardised narrative structures, such as data descriptions and methods, forming a shared framework for interdisciplinary and cross-domain data sharing and reuse. This study shows that while data papers help bridge cognitive and informational gaps between data producers, publishers, and reusers, there is still room for improvement. Existing gaps undermine their role as boundary objects, leading data reusers to rely on databases, journal articles, and other objects. The findings of this study suggest that current data papers bridge the cognitive and informational gaps between data producers, publishers, and reusers, but there is still room for improvement. During the process of scientific communication, the combination and presence of various boundary objects facilitate the flow of information across communities. Another characteristic of data papers as boundary objects is their flexibility. Data events in data papers vary across disciplines, and this study also shows that researchers from different fields have different expectations and needs for data papers. This disciplinary disparity underscores the important role of data papers as boundary objects in cross-disciplinary collaboration, highlighting how the combination of standardisation and flexibility can promote communication and cooperation among researchers from various fields.

Alignment of narrative structure with reusers' needs

Data papers fulfil the foundational requirements of data reusers concerning discovery, filtering, and comprehension. However, during the reuse phase, application-level needs such as trust assessment, accessibility, and reconstruction hints remain inadequately addressed. Data creators and data reusers frequently hold divergent views regarding the essential procedural data (Börjesson et al., 2022). Our analysis also found a mismatch between data papers and the needs of data reusers. For example, reusers expressed a desire for more information on experimental phases, reuse status, and the reputation of the data publication platform. These details are crucial for assessing data reproducibility and reliability, yet they are often not adequately recorded or described in many data papers. The data availability (A) and data publication (DPub) events in the collected data papers only mention where the data is stored, without addressing the reputation of the data repository, which aligns with the findings of Kim (2020). Data reusers often rely on data repository platforms to capture this missing information. In the data reuse process, the meaning-making process centered on data is time-consuming and iterative. During this process, data credibility and usability are prioritised and receive particular attention from researchers, which should also be a core goal of data papers.

Practical implications

The discourse components and data events identified in this study, along with their alignment to the needs of data reusers, offer valuable insights for developing data paper writing guidelines and templates for data journals. These guidelines can enhance the narrative quality of data papers authored by researchers. Huvila and Sinnamon (2022) noted that researchers often lack sufficient tools or methods to effectively communicate their research processes, highlighting the need for better support when sharing data to aid in the documentation and communication of such information. This study reveals that the Material and Method (M) component, which focuses on data collection events, includes critical information such as the people involved, location, time, accuracy, and conditions. The Data Description (De) component, in turn, emphasises raw data records, content attributes, and distribution features. Both components are crucial for data selection and comprehension and are highly valued by data reusers. In designing data paper writing guidelines and templates, attention should be given to presenting data attributes and features from these two dimensions to effectively promote data discovery and utilisation. Furthermore, Data journals should recognise potential disciplinary differences in discourse structures when establishing data policies and submission guidelines, striving to strike a balance between flexibility and standardisation based on the unique characteristics of each field.

This study also underscores the significance of linking data papers with data repository platforms. Data repositories, as platform-mediated solutions, can develop interactive browsing tools that allow users to preview portions of data before downloading. They can incorporate exploratory data analysis techniques, such as visualisation, to facilitate a quick understanding of data distribution features. Additionally, multi-dimensional search interfaces, based on structured data descriptions and indexes, should be designed to enable users to explore and locate data from various perspectives, including time, location, data type, subject, collection method, experimental conditions, and accuracy. By integrating metadata, data papers, journal articles, data blogs, and related videos, repositories can significantly enhance the discoverability and accessibility of data.

Conclusion

In the context of open science, data papers, as a new form of academic publishing, integrate and expand data-centered information and knowledge exchange activities into traditional formal and informal scientific communication. This study examines the alignment between the narrative structure of data papers as boundary objects and the needs of data reusers. Through a dual perspective of textual analysis and reuser interviews, the research finds that data papers serve as bridges between research communities. However, their narrative structure often fails to provide comprehensive, actionable information required by reusers. Therefore, recommendations are made to improve the narrative structure of data papers to better serve the needs of data reusers. As the information environment, intelligent technologies, and research paradigms continue to evolve, the data reuse process on scientific data repository platforms involves various types of subjects or agents, including humans, algorithms, and intelligent agents, bringing revolutionary changes to the publishing, access, processing, and analysis of scientific data. As boundary objects, data papers should be deeply integrated with AI technologies, utilising intelligent data analysis, automated content generation, and optimisation suggestions to enhance their usability and operational effectiveness. This would provide smarter and more personalised support for data reusers and promote collaboration and innovation within academic fields.

Acknowledgements

The authors would like to sincerely acknowledge the valuable feedback and constructive comments provided by the reviewers. This research was supported by the National Social Science Foundation of China (Grant No. 21&ZD334).

About the authors

Cuiyu Qin is a co-first author. She is a PhD student in the School of Information Management at Wuhan University, Wuhan, Hubei, China. Her research interests include data practices, scientific communication, and Human-AI interaction. She can be contacted at qcuiyu@whu.edu.cn.

Qingyu Duan is a co-first author. She is a postdoctoral researcher in the School of Information Management at Wuhan University, Wuhan, Hubei, China. Her research interests include data reuse, information behavior, and digital reading. She can be contacted at Duanqingyu@whu.edu.cn.

Lei Xu is an Associate Professor in the School of Information Management at Wuhan University, Wuhan, Hubei, China. His research interests include knowledge organisation, open data, and digital humanities. He can be contacted at xlei@whu.edu.cn.

Xiaoguang Wang is the corresponding author. He is the Dean of the School of Information Management at Wuhan University, a Professor and PhD supervisor. He is the Director of the Ministry of Education's Philosophy and Social Science Laboratory for Intelligent Computing in Cultural Heritage and the Director of the Digital Humanities Research Center at Wuhan University. His research interests include data management and digital humanities. He can be contacted at wxguang@whu.edu.cn.

References

Borgman, C. L. (2015). Big data, little data, no data: Scholarship in the networked world. MIT press.

Borgman, C. L., & Groth, P. T. (2024). From data creator to data reuser: Distance matters. arXiv preprint arXiv:2402.07926.

Björk, L. (2015). How reproductive is a reproduction?: Digital transmission of text-based documents. Ph.D. thesis, University of Borås, Borås.

Boell, S. K., & Hoof, F. (2015). Using Heider's epistemology of thing and medium for unpacking the conception of documents: Gantt charts and boundary objects. Proceedings from the Document Academy, 2(1), 3.

Börjesson, L., Sköld, O., Friberg, Z., Löwenborg, D., Pálsson, G., & Huvila, I. (2022). Re-purposing excavation database content as paradata: an explorative analysis of paradata identification challenges and opportunities. KULA, 6(3), 1-18.

Chao, T. (2015). Mapping methods metadata for research data. International Journal of Digital Curation, 10(1), 82-94.

Candela, L., Castelli, D., Manghi, P., & Tani, A. (2015). Data journals: A survey. Journal of the Association for Information Science and Technology, 66(9), 1747-1762.

Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., Kirsch, P., Ault, L., Bell, P.J., Bowie, R.C., Leadbetter, A.M., Lowry, R.K., Moncoiffe, G., Harrison, K., Smith-Haddon, B., Weatherby, A., & Wright, D.G. (2012). Making Data a First Class Scientific Output: Data

Citation and Publication by NERC's Environmental Data Centres. Int. J. Digit. Curation, 7, 107-113.

Campbell, H. A., Micheli-Campbell, M. A., & Udyawer, V. (2019). Early career researchers embrace data sharing. Trends in ecology & evolution, 34(2), 95-98.

Chavan, V., & Penev, L. (2011). The data paper: a mechanism to incentivise data publishing in biodiversity science. BMC bioinformatics, 12(Suppl 15), S2.

Das, A. K. (2021). UNESCO recommendation on open science: an upcoming milestone in global science. Science Diplomacy, 39.

Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., & Borgman, C. L. (2011). Science friction: Data, metadata, and collaboration. Social studies of science, 41(5), 667-690.

Frohmann, B. (2004). Deflating information: From science studies to documentation. Toronto: University of Toronto Press.

Faniel, I. M., & Yakel, E. (2017). Practices do not make perfect: Disciplinary data sharing and reuse practices and their implications for repository data curation. Curating research data, volume one: Practical strategies for your digital repository, 1, 103-126.

Gregory, K., Groth, P., Cousijn, H., Scharnhorst, A., & Wyatt, S. (2019). Searching data: a review of observational data retrieval practices in selected disciplines. Journal of the Association for Information Science and Technology, 70(5), 419-432.

Gregory, K., Ninkov, A., Ripp, C., Roblin, E., Peters, I., & Haustein, S. (2023). Tracing data: A survey investigating disciplinary differences in data citation. Quantitative Science Studies, 4(3), 622-649.

Huvila, I. (2011). The politics of boundary objects: hegemonic interventions and the making of a document. Journal of the American Society for information Science and Technology, 62(12), 2528-2539.

Huvila, I. (2022). Improving the usefulness of research data with better paradata. Open Information Science, 6(1), 28-48.

Huvila, I., Andersson, L., & Sköld, O. (2025). Researchers' data processing descriptions—Understanding paradata creation practices and their underpinning instrumentalities. Journal of the Association for Information Science and Technology.

Huvila, I., & Sinnamon, L. (2022). Sharing research design, methods, and process information in and out of academia. Proceedings of the Association for Information Science and Technology, 59(1), 132-144. https://doi.org/10.1002/pra2.611

Kim, J. (2020). An analysis of data paper templates and guidelines: types of contextual information described by data journals. Science Editing, 7(1), 16-23.

Koesten, L., Gregory, K., Groth, P., & Simperl, E. (2021). Talking datasets–understanding data sensemaking behaviours. International journal of human-computer studies,146, 102562.

Li, K., Greenberg, J., & Dunic, J. (2020). Data objects and documenting scientific processes: An analysis of data events in biodiversity data papers. Journal of the Association for Information Science and Technology, 71(2), 172-182.

Li, K., & Jiao, C. (2022). The data paper as a sociolinguistic epistemic object: A content analysis on the rhetorical moves used in data paper abstracts. Journal of the Association for Information Science and Technology, 73(6), 834-846.

Moher, D., Bouter, L., Kleinert, S., Glasziou, P., Sham, M. H., Barbour, V., ... & Dirnagl, U. (2020). The Hong Kong Principles for assessing researchers: Fostering research integrity. PLoS biology, 18(7), e3000737.

Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., ... & Van den Bussche, J. (2011). The open provenance model core specification. Future generation computer systems, 27(6), 743-756.

Nicholas, D., Boukacem-Zeghmouri, C., Rodríguez-Bravo, B., Watkinson, A., Świgon, M., Xu, J., ... & Herman, E. (2018). Early career researchers: Observing how the new wave of researchers is changing the scholarly communications market. Revue française des sciences de l'information et de la communication, (15).

Østerlund, C., & Crowston, K. (2019). Documentation and access to knowledge in online communities: Know your audience and write appropriately? Journal of the Association for Information Science and Technology, 70(6), 619-633.

Penev, L., Chavan, W., Georgiev, T., & Stoev, P. (2012). Data papers as incentives for opening biodiversity data: one year of experience and perspectives for the future. Poster présenté à EU BON: Building the European Biodiversity Observation Network.

Plale, B., & Kouper, I. (2017). The centrality of data: data lifecycle and data pipelines. In Data analytics for intelligent transportation systems (pp. 91-111). Elsevier.

Roa-Martinez, S. M., Vidotti, S. A., & Santana, R. C. (2017). Proposed structure of a data paper structure as scientific publication. Revista Espanola De Documentacion Cientifica, 12.

Swales, J. (1990). Genre Analysis: English in Academic and Research Settings. Cambridge, UK: Cambridge University Press.

Schöpfel, J., Farace, D., Prost, H., & Zane, A. (2019). Data papers as a new form of knowledge organisation in the field of research data. KO KNOWLEDGE ORGANISATION,46(8), 622-638.

Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, translations, and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. Social studies of science, 19(3), 387-420.

Voss, B. L. (2012). Curation as research. A case study in orphaned and underreported archaeological collections. Archaeological Dialogues, 19(2), 145-169.

Wang, X., Duan, Q., & Liang, M. (2021). Understanding the process of data reuse: An extensive review. Journal of the Association for Information Science and Technology, 72(9), 1161-1182.

Yeo, G. (2008). Concepts of Record (2): Prototypes and Boundary Objects. American Archivist, 71(1), 118–143.

Zins, C. (2007). Conceptual approaches for defining data, information, and knowledge. Journal of the American society for information science and technology, 58(4), 479-493.

Zhang, L., Kopak, R., Freund, L., & Rasmussen, E. (2010). A taxonomy of functional units for information use of scholarly journal articles. Proceedings of the American Society for Information Science and Technology, 47(1), 1-10.